EXHIBIT 3

Expert Report of Professor Lee-Jen Wei

2019 WL 2163027 • In Re: TAXOTERE (DOCETAXEL) PRODUCTS LIABILITY LITIGATION. • United States District Court, E.D. Louisiana. (Approx. 11 pages)

4 of 67 results

Original Image of 2019 WL 2163027 (PDF)

2019 WL 2163027 (E.D.La.) (Expert Report and Affidavit)
United States District Court, E.D. Louisiana.

In Re: TAXOTERE (DOCETAXEL) PRODUCTS LIABILITY LITIGATION.

MDL-2740.
No. 216md02740.
February 8, 2019.

Expert Report of Professor Lee-Jen Wei

Name of Expert: Lee-Jen Wei, Ph.D.
Area of Expertise: Accounting & Economics >> Statistics
Representing: Defendant
Jurisdiction: E.D.La.

I. INTRODUCTION

A. Qualifications

1. I received a Ph.D. in Statistics from the University of Wisconsin. I have been a tenured professor of biostatistics at Harvard University since 1991 and was a professor of biostatistical science and computational biology at Dana-Farber Cancer Institute, Harvard Medical School, between 1997 and 2012. I was the scientific director for the Program of Quantitative Sciences for Pharmaceutical Medicine and the co-director of the bioinformatics core at Harvard School of Public Health from 2003 to 2007. From 2003 to 2004, I served as the acting chair of the Department of Biostatistics at Harvard University. I was a tenured full professor of biostatistics and statistics at University of Wisconsin, University of Michigan, and George Washington University from 1982 to 1991.

2. Throughout my career, I have been intimately involved in the design, monitoring, and analysis of clinical studies. I have served on numerous Data and Safety Monitoring Boards for clinical trials and have extensive experience in the evaluation of efficacy and safety data from clinical studies. I have long been actively involved in clinical research and development of a number of novel quantitative methods for analyzing data readily applicable to clinical studies.

3. My scholarly research includes over 185 publications in peer-reviewed journals. I am responsible for developing numerous novel statistical methods for designing, monitoring, and analyzing clinical studies, survival analyses, and meta-analyses. Many of these methods have been included in the most commonly used statistical software packages such as SAS, S-plus, and R. I have served on the editorial boards of a number of statistical journals and am an elected Fellow of the American Statistical Association and Institute of Mathematical Statistics. I was named “Statistician of the Year” in 2007 by the Boston Chapter of the American Statistical Association. In 2009, I received the Wilks Medal from the American Statistical Association, one of the most prestigious awards in the field of statistics, for outstanding contributions to clinical trial methodological research.

4. My *curriculum vitae,* which includes a complete list of my publications, is attached hereto as Appendix A.

Back to top

B. Assignment

5. I was asked by counsel for Sanofi to respond to the question of whether there is reliable statistical evidence that Taxotere is associated with an increased risk of permanent or irreversible alopecia as compared to other cancer-treatment regimens.

6. I was also asked to review and respond to the opinions presented by Dr. Madigan, who filed an expert report on behalf of the Plaintiffs in this matter.

7. All materials reviewed and used for the preparation of my report are listed in Appendix B.

8. A list of my testimony for the last four years is attached as Appendix C.

C. Compensation

9. I am being compensated at the rate of $450 per hour for my time incurred on this matter. My compensation is not contingent on my findings or on the outcome of this litigation.

II. SUMMARY OF OPINIONS

10. The data from the TAX316 and GEICAM9805 studies do not provide any evidence of a safety signal of a new or unexpected risk of permanent or irreversible alopecia with Taxotere use as compared to other cancer-treatment regimens.

11. Dr. Madigan's analyses suffer from serious and well-known limitations, rendering those analyses of little or no value in determining whether there is reliable statistical evidence that Taxotere is associated with an increased risk of permanent or irreversible alopecia.

III. PRIMER ON STATISTICAL ANALYSIS, CLINICAL TRIALS, STUDY ENDPOINT, AND MULTIPLE COMPARISONS

A. Making Inferences About Population Characteristics Using a Sample of Subjects

12. Suppose that we are interested in the rate of occurrence of a certain clinical event (for example, permanent alopecia) among patients treated with Taxotere relative to its counterpart (control) for patients who have been exposed to other treatments. In the first step, we take a sample from a population of patients treated with Taxotere and another sample from the population of patients who did not receive Taxotere. Assuming that these samples are valid representatives of the two populations, quantitative/analytic methods can be used to determine whether the Taxotere group has a higher, lower or similar event rate than that for the non-Taxotere group. Since we draw conclusions based on a subset of patients, any qualitative or quantitative interpretation of the result (i.e., whether the rate is higher or not) is subject to sampling error. That is, the observed event rate may be higher (leading to a possible false positive finding) or lower (leading to a possible false negative finding) than the true event rate in the population. An efficient statistical method for analyzing such data minimizes the chance of making these two types of errors. It is important to note that except for treatment with Taxotere, Taxotere users in the sample ideally should be similar to the patients in the non-Taxotere sample with respect to important observable or unobservable confounders (e.g., age, disease status, et al.), otherwise, adjustment would be needed to make fair comparisons. In other words, it is important that the design of the study “matches” the patients in the two different arms of the study so that the only difference between the two groups is their exposure to the drug and that e.g., age, disease status, are not accounting for differences in the outcomes.

13. After we have determined how to draw a valid sample from the patient population of interest, one has to determine what clinical endpoints are most appropriate to quantify the side effect of the treatment. For the present legal case, the endpoint is whether the patient had permanent alopecia or not. Suppose that based on a sample of 100 patients at the end of study, four patients experienced such events. An obvious estimate of the event rate for the underlying population is 0.04 (or 4%). This is called a point estimate. However, this estimate is based on a sample of patients. The true event rate for the entire population may be more or less than 4%. A different study based on a different sample may find different proportion of patients that experienced alopecia events. Therefore, when

Back to top

observing results from a single sample, it is important to attach a level of confidence to the observed point estimate. This quantitative, scientific process is called “drawing or making inferences” about the true event rate.

14. It is important to report not only the point estimate but also the confidence interval around the point estimate as a measure of the precision of the single point estimate.

15. For a single study with a single endpoint, a commonly used method for inference is to construct a 95% confidence interval for the true event rate. Based on the above sample (four patients had alopecia events among 100 patients), the exact 95% confidence interval for the true rate includes all possible values between the lower bound of 0.011 and the upper bound of 0.099 and is denoted by (0.011, 0.099). That means that if we were to repeat the analysis using different independent samples from the same patient population, say, 1000 times and build 95% confidence intervals around the point estimate each time, 950 of the resulting 1000 confidence intervals would include the true event rate. Since we only have a single dataset and a single confidence interval, loosely speaking, we can say that with 95% certainty, our single observed confidence interval would include the true event rate, which is between 0.011 and 0.099.

16. Now, suppose that we are interested in comparing the event rates across two groups of patients with respect to alopecia events. Suppose that the first group of patients is treated with Taxotere (as discussed previously), and the second group of patients does not take Taxotere (as a control group). Suppose that we observe two events in a sample of 100 patients in the control group and four events in a sample of 100 patients in the Taxotere group. Can we conclude that patients receiving Taxotere have an increased likelihood of experiencing an alopecia event relative to patients who are not taking the drug? To answer this question, we must consider an exact 95% confidence interval for the odds ratio between two true incidence rates (0.04 for the Taxotere group minus 0.02 for the control group) ranging from 0.28 to 22.98. Note that value one (meaning no difference between two groups) is in the middle of the interval. That is, based on these data, we cannot conclude that Taxotere is associated with an increased risk of alopecia events relative to control with a confidence level of 95%. The metric used to compare two groups here is called the odds ratio. As a popular measure in assessing the group difference, the odds ratio is the ratio of two odds. For this example, the odds for Taxotere is 0.042, which is 4% divided by 96% (100% minus 4%) and the odds for control is 0.020, which is 2% divided by 98% (100% minus 2%). The resulting odds ratio for this case is 2.04. An alternative metric to summarize the group difference may be the risk ratio, that is, the ratio of the two rates. For the above example, the ratio would be 4% divided by 2%, which is 2. When the rates are low, the risk ratio is close to odds ratio as in this example. One may also use the risk difference, which is the difference between two incidence rates to summarize the group difference.

17. The confidence interval estimation procedure that I have just described may be used to test hypotheses about the relative incidence rate of alopecia events among patients administered Taxotere relative to patients administered a control. If the 95% confidence interval for the difference between the two rates includes the null hypothesis (“null”) value zero (or if the ratio of the two rates were taken, then the null value will equal one), we cannot conclude that Taxotere is associated with a statistically significantly higher risk with respect to this endpoint relative to control. On the other hand, if the confidence interval does not include zero, a true difference between the two groups may be likely. Since we only have limited data, this claim is subject to error. Here, the probability of making a false positive statement (that is, of stating that there is a difference between the Taxotere and the control group even though there is none) is 5% (the conventional acceptable error probability in the literature for a single endpoint analysis and single study). It is important to note that even if we find that there is a statistically significant difference between the two arms, the next important question is whether that difference is clinically meaningful with respect to risk-benefit considerations.

18. The 95% level for the confidence interval, or the 5% level of significance for hypothesis testing, is typically used by study investigators and statisticians to establish the “statistical significance” of a result when testing a single clinical endpoint in a single study. One may use a 95% confidence interval for the group difference (for a single endpoint and a single study) to assess whether there is a true difference with a 5% error rate. For example, if the true adverse event rates between two groups are the same, then the ratio of the two event rate should be 1 (one). If the 95% confiden interval does not include 1, then we claim that these two event rate are significantly different from each other. The

Back to top

5% level of significance for hypothesis testing can be too "liberal" (i.e., can result in statements of statistical significance when none exists) if multiple endpoints and/or studies are examined simultaneously. In other words, making multiple comparisons inflates the overall "false positive" rate. For example, in the clinical trial investigating the use of Taxotere for treating cancer, the primary endpoint was the efficacy of the drug. However, numerous potential adverse event endpoints also would be examined. Using the 5% rule for claiming statistical significance to analyze simultaneously a large number of safety endpoints in a study will yield a high rate of false positive findings. For example, in the clinical study report for Study TAX316 on Taxotere, there are 69 adverse events reported on Table 7 on pages 37-38. When the 5% test hypothesis rule is applied simultaneously to a number of safety endpoints in a study, the overall false positive rate is as high as 97% (that is, a very high chance of finding at least one endpoint such as alopecia, for which the treatment is not safe, when, in fact, there is no difference between the two groups for all safety endpoints). Inflated false positive rate issues have been extensively discussed, for example, in Friedman, Furberg and DeMets [1] when dealing with multiple testing or simultaneous inferences.

19. A standard procedure to handle the multiple comparison issue is to use the Bonferroni adjustment. For example, if there are 50 different types of adverse events considered in the trial, the total false positive rate is 5%, then for each individual adverse event, we should use a false positive rate of 0.1% (5% divided by 50) to assess whether there is a potential signal on the safety concern. The corresponding confidence interval level should be 99.9% (100% - 0.1%). Note that if we include efficacy endpoints in the comparisons, the threshold value for the false positive rate is even much lower than 0.1%.

20. The problem of inflation of a Type I error (or false positive) [2] rate becomes much worse when we examine the results of several independent clinical trials at the same time with a Type I error rate of 0.05 for each study. For example, suppose there are two independent studies which compare Taxotere with control. Suppose that we claim that there is a significant difference between these two groups when the p-value of any one of these two trials is less than 0.05. If we apply this decision rule, the total Type I error rate would be 9.75%; that is, even if there were no differences between Taxotere and control with respect to alopecia events, the chance of claiming either Taxotere or control is harmful is more than 9.75% in at least one study. This problem is compounded if we apply the same rule to a large number of studies. Therefore, when we analyze multiple studies simultaneously, any conclusion of toxicity has to be carefully interpreted.

B. Using Clinical Trials to Evaluate a Drug Efficacy and Safety

21. Let me turn to the issues of comparing two groups of patients, one receiving Taxotere and the other receiving a control. To make sure that two samples of patients are comparable with respect to all potential confounders, we often rely on a randomized clinical trial setting. Such a clinical study yields a well-designed experiment that has the potential for generating reliable prospective data on drug efficacy or safety. Such studies are conducted and monitored according to a pre-specified protocol which details the treatments administered (e.g., form, dosage, frequency), the clinical or biological endpoints (e.g., lab value, patient's quality of life, time to remission, time to a toxicity event), the study patient population and other clinical and statistical considerations. The trial is usually randomized, which means patients are assigned randomly to one of the study arms. This avoids selection bias or other experimental bias. When appropriately designed, results from a well-conducted, randomized clinical trial are regarded as a gold standard in controlled settings to evaluate the efficacy and safety of a treatment.

22. Suppose that the primary endpoint of a study is the efficacy of Taxotere relative to a control. At the end of the study (or at the interim analysis), a summary statistic is constructed to estimate the relative efficacy of Taxotere and control. That is, the difference between the two groups is quantified. Then, a 95% confidence interval is calculated around this difference. If the confidence interval excludes the "null" value (usually it is zero or one), the two groups are considered statistically significantly different from each other with respect to the primary efficacy endpoint. There are often numerous secondary endpoints including efficacy /safety endpoints in the study. However, because of the multiplicity of comparisons, when "the statistical significance definition" (for example, 5% rule for hypothesis testing) is applied to these secondary endpoints, we usually interpret the results with great caution. That is, if for a specific endpoint, we want to claim that the result is "statistically significant" at the 5% level, we need to use a m

Back to top

lower false positive rate threshold level, for example, 1% instead of 5%, to overcome the multiple comparison testing problem.

23. While clinical trials are often geared towards assessing the efficacy of a particular product, other relevant outcomes are also routinely collected by investigators. Of particular interest is the collection of adverse events and whether or not they are associated with the drug being studied. In fact, the phrase “adverse events” has a broad meaning. The occurrences of these clinical events may not be unusual at all for the general study population. Moreover, such “adverse events” may be related to a host of factors including patient characteristics, associated medical disorders, disease progression, and additional treatments after off the study drug.

IV. ANALYSIS FOR PERMANENT ALOPECIA EVENTS

24. I used safety data from two studies to conduct my analysis-the TAX316 study and the GEICAM9805 study. Before discussing the details of the studies, we must define the term “ongoing” in the context of these two studies. The protocols for each study discuss the need to follow “ongoing clinical adverse experiences possibly or probably related to study drug at the time of End of Chemotherapy.” Whether an adverse event is “ongoing” in these two studies is measured at a fixed point in time at the follow-up visits for the patients. Whether an adverse event is measured again depends on whether there is further follow-up with that patient. It is for this reason that reports of “ongoing” adverse events in these patients do not mean that alopecia was “irreversible” or “permanent.”

25. In addition, defining alopecia as “permanent” or “irreversible” based on the time since chemotherapy, as Dr. Madigan's analysis purports to do, is also inherently unreliable. As used in the adverse event reports and clinical studies, alopecia is recorded or reported as an unwanted side effect. This is why the word “ongoing” is used to describe alopecia. Alopecia in the sense of being “permanent” or “irreversible” would require a medical diagnosis made by a clinician. None of the adverse event reports or clinical study data can appropriately be characterized as cases of “permanent” or “irreversible” alopecia, let alone diagnosed by the clinician as caused by chemotherapy at the exclusion of other causes, without evaluation. [3]

26. The first study, TAX316 (EFC6041/BCIRG001), is “A multicenter phase III randomized trial comparing docetaxel in combination with doxorubicin and cyclophosphamide (TAC) versus 5-fluorouracil in combination with doxorubicin and cyclophosphamide (FAC) as adjuvant treatment of operable breast cancer patients with positive axillary lymph nodes.” The primary goal is to compare disease-free survival (DFS) after treatment with Taxotere (docetaxel) in combination with doxorubicin and cyclophosphamide (TAC) to 5-fluorouracil in combination with doxorubicin and cyclophosphamide (FAC) in operable breast cancer patients with positive axillary lymph nodes. The secondary goals are to compare overall survival (OS), toxicity, and quality of life between the 2 above-mentioned arms, and to evaluate pathologic and molecular markers for predicting efficacy. An interim analysis study report for TAX316 (EFC6041/BCIRG001) was completed on January 21, 2004, and was based on the second interim analysis of TAX316 data, with cut-off dates of 30 April 2003 and 15 July 2003 for safety and efficacy data, respectively.

27. Of the 1491 randomized patients, 11 did not receive any study drugs: 1 in the TAC group and 10 in the FAC group. Eight withdrew consent, 1 was lost to follow-up, and 2 did not receive treatment for other reasons. In total, therefore, 1480 patients were treated with study drugs and are included in the safety analysis (TAC: 744; FAC: 736). One patient randomized to the TAC group (Patient No. 12214) received a combination of docetaxel, doxorubicin, and 5-fluorouracil (TAF) for the first 3 cycles by error followed by 3 cycles of TAC. This patient is analyzed for efficacy and safety in the TAC group. Among the 1480 patients treated with study drugs, 82 patients were lost to follow-up at the end of the study with an actual median follow-up time equal to 96 months.

28. From Table 7 of the final Clinical Study Report, there are 744 and 736 patients in TAC and FAC groups, respectively. For “ongoing” alopecia, there are 29 (4.2%) and 16 (2.5%) cases for TAC and FAC, respectively. [4] The odds ratio (FAC vs. TAC) is 0.55 with a nominal p-value of 0.074. Since there are 69 different types of adverse events reported in Table 7, with Bonferroni multiple comparison adjustment, the Type I error rate “alpha” level should be 0.0007 (0.05 divided by 69). That is, if the p-value is less than 0.0007, then we may claim there is a statistically significant difference between TAC and FAC with respect to ongoing alopecia. Note that the false positive level 0.0007 is for a single stuc

Back to top

there are multiple studies, the error rate would be even smaller. With this threshold value, the level of the confidence interval for the difference between TAC and FAC should be 0.9993. The resulting exact confidence interval of odds ratio (FAC vs. TAC) is (0.165, 1.624), allowing for a potential 60% increase in the alopecia risk associated with FAC compared to TAC.

29. For TAX316, since the p-value is greater than 0.0007, we may not claim that there is a statistically significant difference between TAC and FAC with respect to ongoing alopecia. The data from the TAX316 study therefore does not provide any evidence of a safety signal of a new or unexpected risk of permanent or irreversible alopecia with Taxotere use beyond random chance.

30. The data from the second study is GEICAM9805 (TAX.ES1.301), "A multicenter Phase 3 randomized trial comparing docetaxel in combination with doxorubicin and cyclophosphamide (TAC) versus 5-fluorouracil in combination with doxorubicin and cyclophosphamide (FAC) as adjuvant treatment of high risk operable breast cancer patients with negative axillary lymph nodes."

31. The study initiation date (first patient enrolled): June 21, 1999 and the final report cut-off date (all patients followed for at least 5 years) was March 4, 2009. The follow-up report cut-off date was April 22, 2013. The primary goal of the study was to compare DFS after treatment with TAC to FAC as adjuvant treatment of high risk operable breast cancer patients with negative axillary lymph nodes. The secondary goals were to compare OS between the 2 above mentioned groups, to compare toxicity and QOL between the 2 above mentioned groups, and to evaluate pathologic markers for predicting efficacy. I used the report to compare DFS after treatment with TAC to FAC as adjuvant treatment of high risk operable breast cancer patients with negative axillary lymph nodes after 8 and 10 years and to compare OS between the 2 above mentioned groups after 8 and 10 years.

32. For adverse event analysis, I used Table 47 from the Clinical Study Report dated September 15, 2009. In Table 47, there were 514 and 508 patients for TAC and FAC for alopecia, respectively. There were 3 (6.1%) and 1 (2.9%) "ongoing" alopecia events.[5] The odds ratio (FAC vs. TAC) is 0.34 with a nominal p-value of 0.62 (based on the Fisher's exact test due to the small number of events). In this table, there were 40 different types of adverse events reported. Using the Bonferroni multiple adjustment, the Type I error rate (false positive rate) should be 0.00125. That is, if the p-value is less than 0.00125, we may claim that there is a statistically significant difference between TAC and FAC with respect to ongoing alopecia.

33. For GEICAM9805, since the p-value is greater than 0.001, we may not claim that there is a statistically significant difference between TAC and FAC with respect to ongoing alopecia. This is corroborated by the exact 99.93% confidence interval of the odds ratio (FAC vs TAC): (0.00009, 21.56), which includes the null value one. The data from the GEICAM9805 study therefore does not provide any evidence of a safety signal of a new or unexpected risk of permanent or irreversible alopecia with Taxotere use beyond random chance.

34. From the data from these two studies, there is no evidence that TAC would increase the incidence of "ongoing," "permanent" or "irreversible" alopecia versus its comparators.

V. REVIEW OF DR. MADIGAN'S ANALYSES

A. Dr. Madigan's FAERS Analysis

35. I have reviewed the FAERS analysis Dr. Madigan conducted. That analysis suffers from a number of well-known limitations.

36. As Dr. Madigan himself has acknowledged, spontaneous reporting system (SRS) data cannot be used to definitively establish a cause-and-effect relationship: "...SRS data can never be used to definitely establish cause-and-effect relationships..."[6]

37. Dr. Madigan reiterated in 2013 that disproportionality analyses (DPA) methods like the ones he uses here are only hypothesis generating when applied to spontaneous reporting databases:

Back to top

In the context of spontaneous report systems, some authors use the term "signal of disproportionate reporting" (SDR) when discussing associations highlighted by DPA methods [6, 7]. In reality, most SDRs that emerge from spontaneous report databases represent non-causal effects because the reports are associated with treatment indications (i.e., confounding by indication), co-prescribing patterns, co-morbid illnesses, protopathic bias, channeling bias, or other reporting artifacts, or, the reported adverse events are already labeled or are medically trivial. In this sense, SDRs generate hypotheses. [7]

38. In addition to Dr. Madigan's own publications, other scientists have recognized that the main usage of SRS data is for initial exploration and information mining, not to establish a causal relationship between a drug exposure and an adverse event. [8]

39. Another limitation of Dr. Madigan's FAERS analysis is that it does not take into account stimulated reporting. Dr. Madigan's disproportionality analysis essentially compares the reporting rate of “irreversible alopecia” adverse events for docetaxel to the reporting rate of “irreversible alopecia” adverse events for all other drugs. The principle behind this comparison is that if “irreversible alopecia” is reported more frequently for docetaxel than for other drugs it may signal a possible association that needs to be investigated. One underlying assumption to that analysis is that all things are otherwise equal. This assumption may not be true in the case of docetaxel.

40. Unlike the reporting for all other drugs in the FDA database, reporting of “irreversible alopecia” events for docetaxel could have been stimulated by a number of things, including: (1) media reporting about docetaxel and alopecia and (2) the 2015 change to the Taxotere label involving permanent alopecia/hair loss. FDA itself has recognized that FAERS data “may be affected by the submission of incomplete or duplicate reports, under-reporting, or reporting stimulated by publicity or litigation. As reporting biases may differ by product and change over time, and could change differently for different events, it is not possible to predict their impact on data mining scores.” [9] Although Dr. Madigan claims to exclude lawyer reports, his analysis does not, and cannot, account for these factors. For example, Dr. Madigan cannot rule out that the increase in docetaxel reports he found in 2016 and 2017 is the result of reporting stimulated by publicity or litigation.

41. Dr. Madigan's attempt to compare the proportional reporting rates of docetaxel with that of other select chemotherapies suffers from the same methodological limitations. Dr. Madigan included the other select chemotherapies in his analysis and graphs so that he and the other Plaintiffs' experts who rely on his work can argue that docetaxel causes “irreversible alopecia” and the other select chemotherapies do not. FDA has recognized the limitations of this kind of work, and has repeatedly stated that comparisons of the reporting rates of different drugs are hypothesis generating and must be viewed and undertaken with extreme caution:

> Comparisons of reporting rates and their temporal trends can be valuable, particularly across similar products or across different product classes prescribed for the same indication. However, such comparisons are subject to substantial limitations in interpretation because of the inherent uncertainties in the numerator and denominator used. As a result, FDA suggests that a comparison of two or more reporting rates be viewed with extreme caution and generally considered exploratory or hypothesis-generating. Reporting rates can by no means be considered incidence rates, for either absolute or comparative purposes. [10]

42. Dr. Madigan's FAERS analysis suffers from other well-known limitations. For example, database studies like the one Dr. Madigan performed here often lack information on confounding factors [11] and often include inaccurate or incomplete data. [12]

43. Additionally, the definition of the reporting ratio (RR), the parameter in Dr. Madigan's analysis, is not a direct group contrast measure we can easily interpret. For example, consider the adverse event “nausea” observed from patients treated by drug G. Conventionally we would be interested in comparing two probabilities:

pr(nausea | drug G) and pr(nausea).

However, we cannot use the SRS dataset to estimate these two quantities because we do not have enough information about the population with or without exposure from drug G. Instead, we only have information on the population of patients, who had or had no exposure to drug G but with certain types of signals or features to be selected. For example, one restriction is that a patient has to report some type of adverse events (AEs) to be inclu

Back to top

in the SRS data population. Therefore, the RR used in the current report is equivalent to a contrast between two probabilities:

pr(nausea | drug G ± reported AE) and pr(nausea | reported AE).

If the probability of nausea among patients taking drug G and reporting some AEs is different from the probability of nausea among patients taking drug G, we cannot obtain an unbiased assessment on the drug safety in the presence of such a selection bias.

44. We can use the example on page 6 of Dr. Madigan's report to demonstrate this issue. Suppose that there are 5,400 patients without experiencing any AE (including Nausea) not included in the SRS databases. Among them, 1,080 patients actually took Ganclex. In this case, the risk of Nausea among patients taking Ganclex is 20/1200=1.67% and the risk of Nausea among the entire population is 120/6600=1.82%. While the true risk of Nausea among Ganclex users is lower than the general population, the reporting ratio (RR) based SRS databases is 1.67, suggesting that the risk of Nausea is higher among Ganclex users. This erroneous conclusion based on RR is due to the fact that 5,400 patients not in the SRS database are not included in the analyses. Nausea Nausea Nausea Nausea Total

TABULAR OR GRAPHIC MATERIAL SET FORTH AT THIS POINT IS NOT DISPLAYABLE

45. The empirical Bayes (EB) strategy is the main analytic tool for Dr. Madigan's analysis. The results depend on and are sensitive to a set of strong model assumptions including the specification of the prior distribution and the Poisson assumption for the distribution of the cell counts. Unfortunately, such model assumptions are selected for mathematical convenience and not “tested” empirically since they are nearly not identifiable. Dr. Madigan's report also didn't provide details on which pairs of drug and AE are included in the EB analysis, which also has a significant impact on his result.

46. Moreover, for the same reason, the utilization of a lasso-regularized logistic regression to adjust the so-called “innocent bystander” effect is not robust. [13] Whether the confounding effect can be appropriately controlled depends on the validity of the model specification. This approach puts very strong assumptions (again, cannot be empirically tested for its validity) for the effect on different drug exposures. For example, it assumes that if Drugs A and B increase the risk of an adverse event from 10% to 15% and 20%, respectively, then taking both drugs together would increase the risk from 10%- 28.4%, a magic number based on the logistic regression without interactions. Different model specifications would result in completely different results.

47. Furthermore, lasso-regularized logistic regression assumes a particular prior distribution for the regression coefficients in the “working” logistic regression model. This step is needed, since without appropriate prior distribution, the logistic regression cannot be fitted well with the limitation of the sample size in SRS. However, there is no empirical evidence on the appropriateness of the choice of this prior distribution, which statisticians mainly use for computational convenience. In addition, the lasso regularization is in particular not a satisfactory tool for adjusting the “bystander” effect, since it is well known that “if there is a group of variables among which the pairwise correlations are very high, then the lasso tends to select only one variable from the group and does not care which one is selected.” [14]

48. In other words, when two drugs tend to be used together and are associated with a higher risk of an adverse event, then the lasso-regularization may “randomly” select one drug as the responsible one.

49. Furthermore, Table 3 in Dr. Madigan's report only provides the estimated odds ratios (ORs), but not the corresponding credible or confidence intervals. Without an interval estimate for the “true” odds ratio, it is not clear how reliable the reported estimated ORs are. The choices of various specifications of the model maximize the exploratory capacity but may introduce bias from strong incentive to believe in particular outcome. [15] Experts in the field recommend taking a more skeptical view and cross-examining the results with those generated from other models. [16]

50. The concern of inflated type error due to multiple testing would require a substantially more stringent criterion than SEB05 used by Dr. Madigan for safety signal detection, even if we accepted the current SRS-based analysis. D Back to top Madigan's report (paragraph 38) claims that “Figure 3 shows a safety signal for docetaxel by the middle of 2012.” This

claim is based on the simple observation that SEB05 of Taxotere crosses the threshold level 2.0 (a subjective choice without a solid scientific reason) at the second quarter of 2012. However, this is the 50th test based on Dr. Madigan's analysis. Although a SEB05 above 2.0 ensures that the underlying RR is greater than 2.0 with a 90% confidence at a single time point, continuously monitoring based on SEB05 curve would greatly reduce the reliability of such a claim since the error rate from individual test would accumulate over time. For example, the total false positive error rate, i.e., the probability of false safety alarm, from 50 independent examinations would be above 99%, even if we control the false positive error rate of each individual test at the level of 10%. In the current case, the 50 tests (up to 2012) are not independent, but the total false positive error rate based on the unadjusted SEB05 would still be substantially higher than 10%. Appropriate adjustment for multiple testing requires more stringent criterion. [17]

51. For these reasons, Dr. Madigan's FAERS analysis does not provide reliable statistical evidence that Taxotere is associated with an increased risk of permanent or irreversible alopecia.

B. Dr. Madigan's Analysis of GEICAM9805 and TAX316

52. I also reviewed Dr. Madigan's analysis of irreversible alopecia in the GEICAM9805 (TAX301) and TAX316 studies.

53. Dr. Madigan conducted comparisons between TAC and FAC with the alopecia data from TAX316. The risk ratio (FAC vs. TAC) is 0.55 with 95% confidence interval (0.31, 1.02), which include numerical value of one. As discussed earlier in my report, this means that there is no statistically significant difference between two groups even with the nominal confidence level of 95%. Since there were numerous efficacy and safety endpoints considered in the study, we need to make adjustment of choosing appropriate confidence level and false positive rate to ensure whether there was a real safety signal or not. For instance, if we only consider safety endpoints (not including efficacy endpoints), the false positive rate for individual test should be controlled at the level of 0.0007, not 0.05 for claiming a statistically significant group difference for the present case. Correspondingly the confidence level for constructing confidence interval should be 0.9993, not 0.95.

54. In paragraph 52 of his report, Dr. Madigan considered the elapsed time between the end of the treatment (plus 30 days) and any alopecia resolution (see Table 5 in his report). Since the case is “permanent” or “irreversible” alopecia, not resolution time analysis, the results are not relevant. Furthermore, the p-values reported in Table 5 are nominal values. Since the data were repeated analyzed sequentially over time, it is known that those nominal p-values are not correctly quantifying the group difference.

55. Finally, Dr. Madigan conducted a random effects meta-analysis combining the data from the two studies at completion yields a rate ratio of 1.85 with a corresponding 95% confidence interval (1.04, 3.31) and a p-value of 0.04. This analysis has a serious flaw. In estimating a group difference with the data from multiple studies, meta-analysis has often been misused to combine information across the studies. The choice of the individual studies for meta-analysis is generally ad hoc and can be biased due to the subjectivity of the researcher. The patient populations can be quite heterogeneous across selected studies. The conventional statistical procedures for combining information from multiple studies are mainly based on the fixed or random effects models. [18], [19], [20], [21] The fixed effect model assumes that the true treatment effects are approximately constant across the studies. [22] This assumption is rarely valid in practice and the standard goodness of fit tests for this assumption usually do not have enough statistical power to either refute or support this claim. [23] On the other hand, under the random-effects model, one assumes that the individual studies were random samples from a hypothetical “super-population” of studies and that the true group difference may differ from one study to another but follows a specific distribution such as a normal (or a transformation thereof) distribution across the studies. The combined estimate for the overall treatment effect is then a weighted average of observed study-specific treatment effects where the weights depend on the data.

56. The random effects modeling approach used by Dr. Madigan requires two assumptions [24], [25]: (1) the number of studies needs to be large; and (2) the estimated risk ratio from each individual study needs to follow a normal distribution. In the current analysis, there are only two studies and it is impossible to reliably estimate the distribution of the super-population (for instance, we cannot estimate the mean and variance of a normal distribution based on two observations with reasonable accuracy). In addition, there are only three alopecia ever and one alopecia event in two treatment arms of GEICAM9805. Thus, the risk ratio estimate in study GEICAM9805

Back to top

does not follow a normal distribution. Therefore, the reported 95% confidence interval (1.04, 3.31) and p-value in Dr. Madigan's meta-analysis are not valid and cannot be used as statistically significant evidence for the presence of any alopecia risk. That is, the confidence interval estimate for the treatment effect can be substantially smaller than necessary to reflect the true uncertainty in the data. Moreover, the significant threshold value (Type I error rate) for alopecia event comparison is much smaller than 0.05 due to multiple comparison issue discussed previously in this report.

57. In conclusion, based on the results from my analysis and the flawed arguments from Dr. Madigan, I cannot find any statistical evidence that shows Taxotere increases the risk of permanent or irreversible alopecia as compared to other cancer-treatment regimens.

Footnotes

[1] For details see Friedman, Furberg and DeMets, Fundamentals of Clinical Trials, Second Edition, Chapter 15; p. 215, Littleton, MA, 1985.

[2] Type I error rate is defined as the probability that the test will reject a true null hypothesis.

[3] V Sibaud, et al., Dermatological adverse events with taxane chemotherapy, Eur. J. Dermatol. (2016) Oct. 1; 26(5); 427-443.

[4] If you analyze the entire safety population, which would be all patients that received TAC or FAC, the percentages of ongoing alopecia are 3.9% in the TAC arm and 2.2% in the FAC arm.

[5] If you analyze the entire safety population, which would be all patients who received TAC or FAC, the percentages of ongoing alopecia are 0.6% in the TAC arm and 0.2% in the FAC arm.

[6] M Hauben, D Madigan, C Gerrits, L Walsh and E Puijenbroek, The role of data mining in pharmacovigilance. Expert Opin. Drug Saf. (2005) 4(5):929-948.

[7] W DuMouchel, P Ryan, M Scheumie, D Madigan, Evaluation of Disproportionality Safety Signaling Applied to Healthcare Databases. Drug Saf. (Oct. 2013) 36.1.

[8] B Strom, Evaluation of suspected adverse drug reactions. JAMA (2005) 293(11): 1324-1325.

[9] FDA Guidance: Good Pharmacovigilance Practices and Pharmacoepidemiologic Assessment (Mar. 2005) ("FDA Guidance").

[10] FDA Guidance at 11.

[11] W Ray, Improving automated database studies. Epidemiology (2011) 22(3); M Hauben, D Madigan, C Gerrits, L Walsh and E Puijenbroek, The role of data mining in pharmacovigilance. Expert Opin. Drug Saf. (2005) 4(5):929-948.

[12] P Coloma, M Schuemie, G Trifir6, R Gini, R Herings, J Hippisley-Cox, G Mazzaglia, C Giaquinto, G Corrao, L Pedersen, et al. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. Pharmacoepidemiol. Drug Saf. (2011) 20: 1-11; M Hauben, D Madigan, C Gerrits, L Walsh and E Puijenbroek, The role of data mining in pharmacovigilance. Expert Opin. Drug Saf. (2005) 4(5):929-948.

[13] M Hauben, D Madigan, C Gerrits, L Walsh and E Puijenbroek, The role of data mining in pharmacovigilance. Expert Opin. Drug Saf. (2005) 4(5):929-948.

[14] H Zou and T Hastie, Regularization and variable selection via the elastic net. J. R. Statistit. Soc. (2005) (B) 67:301-320.

[15] M Hauben, L Reich, Application of an empiric Bayesian data mining algorithm to reports of pancreatitis associated with atypical antipsychotics. Pharmacother. (2004) 24(9):1122-1129.

[16] T Louis, W Shen, Discussion for "Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system." The American Statistician (1999) 53(3); D Madigan Discussion: The American Statistician (1999) 53(3).

[17] I Silva and M Kulldorff, Continuous versus group sequential analysis for post-market drug and vaccine safety surveillance. Biometrics (2015) 71(3): 851-858; I Silva, Type I error probability spending for post-market drug and vaccine safety surveillance with Poisson data. Methodol Comput Appl. Probab. (2018) 20:739-750.

Back to top

18 KR Abrams, DR Jones, TA Sheldon, F Song. Methods for meta-analysis in medical research. New York: J Wiley; 2000.

19 SLT Normand. Tutorial in biostatistics meta-analysis: formulating, evaluating, combining, and reporting. Statistics in Medicine (1999) 18:321-359.

20 LV Hedges, I Olkin. Statistical methods for meta-analysis. New York: Academic press; 1985.

21 R DerSimonian, R Kacker. Random-effects model for meta-analysis of clinical trials: an update. Contemporary Clinical Trials (2007) 28: 105-114.

22 JPT Higgins, SG Thompson, DJ Spiegelhalter. A re-evaluation of random-effects meta-analysis. Journal of the Royal Statistical Society A (2009) 172:137-159.

23 JPT Higgins, SG Thompson. Quantifying heterogeneity in a meta-analysis. Statistics in Medicine (2002) 21:1539-1558.

24 SE Brockwell, IR Gordon. A comparison of statistical methods for meta-analysis. Statistics in Medicine (2001) 20:825-840.

25 M Henmi, JB Copas. Confidence intervals for random effects meta-analysis and robustness to publication bias. Statistics in Medicine (2010) 29:2969-2983.

End of Document